Creating a corpus of sensitive and hard-to-access texts: Methodological challenges and ethical concerns in the building of the WiSP corpus

Journal Item

Applied Corpus Linguistics 1 (2021) 100011

Contents lists available at ScienceDirect

Applied Corpus Linguistics

journal homepage: www.elsevier.com/locate/acorp

Creating a corpus of sensitive and hard-to-access texts: Methodological challenges and ethical concerns in the building of the WiSP Corpus

Maria Leedham*, Theresa Lillis, Alison Twiner

The Open University, School of Languages and Applied Linguistics, Faculty of Wellbeing, Education and Language Studies, Level 1, Stuart Hall Building, Walton Hall, Milton Keynes MK7 6AA, United Kingdom

ARTICLE INFO

Keywords:
Archiving
Anonymisation
Data sharing
Social work
Writing

ABSTRACT

Corpus linguistics is increasingly employed to explore large, publicly-available datasets such as newspaper texts, government speeches and online fora. However, comparatively few corpora exist where the subject matter concerns sensitive topics about living individuals since, due to their highly personal and confidential nature, these texts are hard to access and raise difficult ethical questions around secondary data analysis. One exception is the Writing in professional social work practice (WiSP) corpus, comprising texts written by UK-based professional social workers in the course of their daily work and now available to other researchers through the ReShare archive. This paper focuses on the challenges involved in building the WiSP corpus and the epistemological and ethical issues raised. Two key aspects of research practice are discussed: data anonymisation and dataset archiving. Specifically, the paper explores decision-making around anonymisation and an ethically-informed rationale for treating some texts as 'not for sharing', leading to the decision to create two corpora: one for the research team and a further anonymised and slightly reduced version for archiving. The paper explores what the WiSP corpora (Corpus 1 and Corpus 2) contribute to understandings about social work writing, the extent to which the two corpora enable different analyses and whether the existence of two corpora is problematic from a corpus linguistic perspective. The paper concludes by considering how the ethical decisions around corpus creation of sensitive texts raise questions about key principles in corpus linguistics.

1. Introduction

This paper draws on our experience of working with the UK-based WiSP corpus dataset (Lillis, Leedham and Twiner, 2019), exploring how the creation of a hard-to-access corpus of sensitive texts raises challenging methodological issues in relation to corpus compilation and the additional preparation required to meet the funders' requirements for secondary archiving. Given that social work involves engagement with people at vulnerable points in their lives, the texts produced are often sensitive and always highly confidential, thus making access understandably difficult. One means of resolving some of the ethical issues is the creation of a separate corpus for archiving with the more sensitive texts withheld. While an effective solution to the risks inherent in archiving texts about living, vulnerable people, this response raises epistemological questions about fundamental and widely accepted principles governing corpus linguistics.

Sensitivity of topic and inaccessibility go together as, by their nature, texts concerning sensitive subjects and giving information on specific individuals are not available in the public domain. Within corpus linguistic research in general, and corpus-assisted discourse studies in particular, due to the quest for ever-larger datasets and the ease with which internet-trawling can provide bespoke corpora, less accessible texts have been little-studied (Lischinsky 2018; Marchi and Taylor 2018; Taylor and Marchi 2018). Conventionally-accepted principles of corpus building rest on the foundational quest for representativeness across the whole population of texts under investigation, and assume researcher access to a large number of texts fitting a previously-generated sampling frame, with limitations placed on wordcounts and numbers of texts within particular domains, genres and by individuals (e.g. Sinclair 2005; Koester 2010; McEnery and Hardie 2012). However, in cases where paucity of available data or extreme sensitivity severely limit access to texts, what are the options available to the researcher? In this paper, we use the compilation of the WiSP corpus as our framing for examining these principles and explore the ethical considerations and ultimately the solutions we came to in order to make the corpus more widely available.

The paper discusses the methodological, ethical and epistemological considerations around anonymising a relatively small corpus (1 million words) of sensitive texts, considering institutional access issues, our values and commitments as researchers to participants (including social

* Corresponding author.
E-mail addresses: Maria.leedham@open.ac.uk (M. Leedham), Theresa.lillis@open.ac.uk (T. Lillis), Alison.twiner@open.ac.uk (A. Twiner).

https://doi.org/10.1016/j.acorp.2021.100011
Received 15 July 2021; Received in revised form 18 October 2021; Accepted 2 November 2021
Available online 5 November 2021

workers, service users and third parties such as health visitors or clients' family members) and the imperative to archive the resulting datasets. One significant solution adopted for WiSP was the creation of two corpora: 'WiSP Corpus 1′ for use by the research team and 'WiSP Corpus 2′ for deposit in the grant funders' repository: ESRC ReShare archive.

Research questions addressed in this paper are:

1) What are the challenges in data preparation for archiving hard-to-access, sensitive textual data, particularly around anonymisation coding?
2) Given the adopted solution of creating a separate corpus for archiving, how do the differences between the two versions of the WiSP corpus affect findings?
3) To what extent do the issues raised and solutions found problematise conventionally-accepted principles in corpus building?

Section 2 sets out the decision-making involved in the compilation of the WiSP corpus, in the context of concerns about preserving anonymity and degrees of access of corpora. In Section 3, we consider the ethical challenges and solutions of creating the two corpora, with discussion of anonymisation coding and archiving the corpus. We then explore the extent to which differences between the two versions of the WiSP corpus affect findings, using three worked examples (Section 4). Section 5 underlines the implications arising from the building of the WiSP corpus for corpus linguistics centreing on sensitive texts.

2. Corpus compilation of sensitive and hard-to-access texts

In this section we set out some conventional principles for building a corpus (2.1) before discussing the shortage of corpora comprising sensitive and hard-to-access texts (2.2). Using this framing we then describe the process of designing and building the WiSP corpus (2.3).

2.1. Conventional principles of corpus-building

Before exploring the lack of corpora containing sensitive texts, it is first useful to bring to the fore relevant and widely-accepted principles of corpus compilation. The principles below draw heavily on Sinclair (2005) as a significant and influential figure in the field and are expressed in terms of written rather than spoken corpora (since the WiSP corpus contains only written texts). The core assumption underlying these principles is that the total population of texts is too large to include in its entirety. The four identified principles are:-

1) The corpus should comprise a representative sample of the population of texts under investigation

Since it is generally not possible to include all texts under investigation, it is necessary to select texts 'on the basis of stringent criteria or by random sampling' (Mautner, 2019:8). A clear difficulty here is in determining the nature of the totality of texts since we may only discover that a corpus is unrepresentative if the results are skewed (cf. Tognini-Bonelli 2001). Many means of sampling rely on all items within the target population being available for selection (see discussion in Biber, 1993).

Principles 2 and 3 each relate to this over-riding aim.

2) The larger the corpus, the more representative

The principle that a larger corpus is more representative than a smaller one is often taken as a truism. But what counts as 'large' or 'small' in corpus terms? Three decades ago, Sinclair (1991:18), suggested that 'a corpus should be as large as possible, and should keep on growing'. While Sinclair was discussing monitor corpora (i.e. corpora which are continually added to), in general it is widely regarded that more is better since more data means more instances of the linguistic phenomena under investigation and a reduced chance of skewed results from outliers. However, quantity alone is, of course, no guarantee of representativeness (Lischinsky, 2018; Mautner, 2019), and a sample should ideally contain evenly-distributed examples of the dimensions across which the population varies.

3) The wordcount and number of texts from any individual topic area, genre or writer should be carefully controlled

The ideal for corpus compilation is to map out the 'textual universe' and then construct an a priori sampling frame with each cell populated with equal wordcounts and numbers of texts (McEnery and Hardie, 2012). In addition, restricting texts from individual writers ensures idiosyncratic uses of language do not dominate findings.

4) Any removal of information from the corpus should be limited, as this results in loss of data and authenticity

This principle relates to Sinclair's (1991) preference for the whole text in order to preserve authenticity. While texts may be augmented through headers to provide additional information, any removal of data should be done sparingly (e.g. visuals may be removed and replaced with tags, or personal data such as names or birthdates may be replaced with codes).

Overall, and as has frequently been pointed out, while full representativeness may be unattainable (e.g. Leech, 1992; McEnery and Hardie, 2012), seeking to construct a *balanced* corpus is a viable approximation of this (e.g. Leech, 2007, in Lischinsky, 2018). Crucial to the endeavour is transparency - making clear how the corpus was compiled and the ways in which it may be biased - alongside careful wording of claims to ensure that findings are based on an (of necessity) imperfectly-balanced dataset. A key argument is that the more confidence there is in the representativeness of a corpus, the greater validity there is in making generalisable claims based on the corpus (Mautner, 2019).

Many corpora today are larger than ever before with multi-billion-word collections of texts readily compiled from webscraping publicly-available internet sites (e.g. the Oxford English Corpus) or through accessing large databases (e.g. corpora based on Lexis Nexis). Building corpora from large, readily-available sources of textual data means that the issue of access to such data is unproblematic, and decisions in corpus compilation entail questions such as how big should the sample be? what sampling frame should be applied? what tagging is needed?

In contrast, corpora built from sensitive, hard-to-access texts necessitate questions such as: what texts are available to us? how should we anonymise the texts? will any individual be harmed through the creation of this corpus? Due to the difficulties in assembling and preparing texts, such corpora are 'small' in corpus terms, but are valuable in enabling insights into occluded texts and genres (Swales 2004). Extensive and critical consideration is required at all stages of designing and building such a corpus. Section 2.2 explores the nature of and consequently the lack of such corpora.

2.2. Lack of corpora containing sensitive and hard-to-access texts

Sensitive data includes but goes beyond the legal definition of 'personal data'. Personal data comprises direct and indirect identifiers, with the former referring to 'attributes or combinations of attributes that are structurally unique for all persons in your data' (Elliot et al., 2020:48) such as an individual's name or NHS number (UK Government 2018). Examples of indirect identifiers include a combination of age, birthplace and marital status, meaning that someone may be identified through a combination of rare variables. Personal information may be considered sensitive where there is a risk of damage to the individual if the information is misused or mishandled. Whilst there is general agreement that information on health or criminal convictions always comprises sensitive data, Simitis (1999:5, cited in Elliot et al., 2016), suggests that any item of personal data can be sensitive 'depending on the purpose or the circumstances of the processing'. For example, natal sex may be sensitive data where someone has transitioned (Staples et al., 2018).

As already identified, sensitivity of topic and inaccessibility go together as, by their nature, texts concerning sensitive topics are not avail-

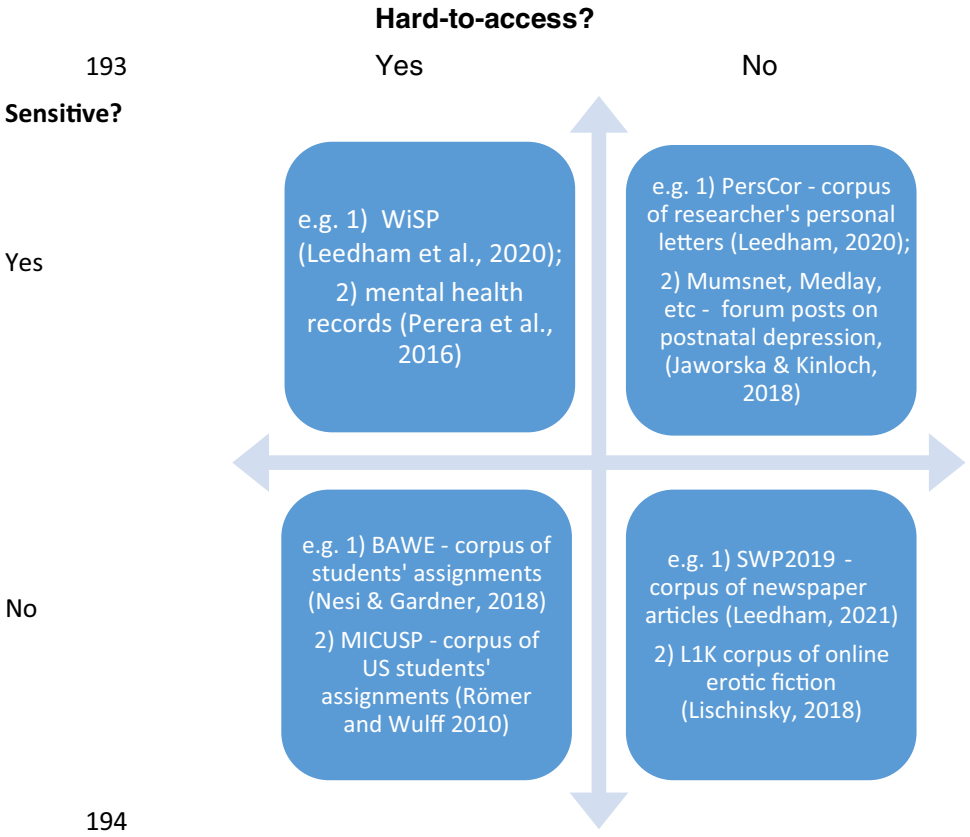

Fig. 1. Examples of corpora in each category.

able within the public domain. Few corpora exist which both contain sensitive texts and are hard-to-access, leading to a situation in which 'corpus linguists who engage in discourse analysis privilege certain text types or registers, at the expense of others' (Baker 2018:283). To provide further evidence of this point, we reviewed all studies in two key journals (*Corpora* and *Journal of Corpora and Discourse Studies*) over a 4-year period (2017–2020 inclusive). Findings indicate that corpus studies overwhelmingly employ publicly-available texts and, whilst some corpora focus on socially and personally sensitive topics (e.g. abortion in Kreischer 2019), few of the texts in such corpora include personal, potentially identifying information which give rise to difficult questions about the ethical concepts of consent, anonymity and confidentiality. Where texts involving sensitive data are included in a corpus (e.g. Ohashi et al., 2020; Bolly and Boutet 2018), there tends to be little discussion of the ethical principles governing anonymisation procedures or the practical challenges faced and decisions taken. A few large corpora of sensitive and hard-to-access texts do exist - mainly within medical research - but access is restricted to a carefully-defined group of researchers who provide aggregated data to others on request (e.g. Perera et al., 2016). Fig. 1 provides a matrix of categories of sensitive and hard-to-access corpora.

The top right quadrant is exemplified through a corpus of one author's personal letters as this is easy for an individual author/analyst to obtain but may be sensitive in subject matter (Leedham, 2020). In the lower right quadrant, corpora built from a newspaper archive such as Lexis Nexis are neither sensitive nor hard-to-access as these texts are in the public domain (e.g. the Social Work Press corpus 2019 [SWP2019], (Leedham, 2021). The lower left quadrant names the British Academic Written English (BAWE) corpus (Nesi and Gardner 2018) to exemplify the category of texts which are occluded (since they are not in the public domain) yet not deemed sensitive (personal information such as student name/number, tutor name, university was immediately anonymised and original data not stored). Finally, the top left shows the area occupied by the WiSP corpus: hard-to-access texts which are also sensitive.

2.3. On the shift towards archiving

In recent years there has been a move towards greater archiving of datasets in national or international repositories (Tilley and Woodthorpe 2011; Irwin 2013; Parry and Mauthner 2004). Archiving data in this way (i.e. for use by researchers beyond a project or single institution) is frequently *required* by research funding bodies and *highly recommended* by journals. In addition to reducing costly repetition of data collection, an important rationale for archiving datasets is to enable greater access to datasets, thus encouraging replication of studies and verification of findings.

This shift to archiving datasets greatly affects the collection and treatment of data, as highlighted by Carusi and Jirotki (2009:287): 'Gathering and storing data for one's own use or with an eye to other possible addressees besides oneself are two very different activities.' The ethical risks in the different types of data are made clear: 'whereas quantitative data are arrived at through abstraction from a context, qualitative data is highly contextualised.' (op.cit., p.38). Following this definition, texts in a corpus constitute qualitative data as they comprise discursive, contextualised natural language, rather than abstractions (cf. Hayes and Devaney 2004). Where such texts contain extensive writer metadata and/or are linked to further datasets, the text is more richly-contextualised and it may prove possible to identify individuals featured in the text through a process of piecing together the 'jigsaw' of information from different sources (http://www.transparencyproject.org.uk/jigsaw-identification).

The ESRC ReShare archive (https://reshare.ukdataservice.ac.uk/) features three levels in which data collections can be deposited: open access (anyone can download), 'safeguarded' (data users must be registered) and 'controlled' (unavailable, though potential users can con-

tact the data controller). Searching in the ESRC ReShare archive indicates that of the 7000+ data collections, just 31 are tagged as a corpus (search conducted in April 2021). Of these, 12 are open access (including BAWE), 15 are 'safeguarded', of which just two require permission from the data controller (one of these is WiSP). A further three are not available 'due to legal, ethical or commercial contracts' with only metadata and documentation such as interview protocols and descriptions of the method deposited in the archive (allowed by funders under particular circumstances).

The issue of what and how to anonymise sensitive data and how to archive the resulting corpus comprises a significant ethical challenge and is discussed in Section 3, drawing on WiSP to exemplify the issues.

3. Creating the WiSP corpus: ethical challenges and solutions

3.1. Introducing the WiSP corpus

The WiSP corpus is a 1-million-word collection of over 4600 texts produced by 38 social workers within three UK local authorities from 2015 to 2017. The corpus was compiled as part of the WiSP[1] project. This project focuses on five local authorities (LAs) in the UK, exploring both the range of written texts produced and the writing practices of social workers. The WiSP project has an overarching ethnographic orientation (see Lillis, Leedham and Twiner, 2017, 2020), using multiple methods of data collection and analysis (e.g. 10 weeks observations, 81 interviews, 29 social worker writing logs, institutional documentation) to explore how writing is situated within social workers' daily working lives. While we are aware that ethical regulations for research vary across different countries, we hope that our description of the WiSP project is useful to all researchers.

The aim in designing the WiSP corpus was to produce a 'snapshot' corpus which, to use Hardie and McEnery's definition (2019:9), seeks 'balance and representativeness within a given sampling frame', thus following the four principles outlined in 2.1. The first challenge in compiling the WiSP corpus was negotiating access: Ethical approvals were required from our own institution (The Open University Human Research Ethics Committee) and from each LA as well as a UK enhanced Disclosure and Barring Service (DBS) checks for all researchers and an additional data sharing agreement with one LA (for full details see Lillis, Leedham and Twiner, 2017). In seeking access to texts and in social worker participation, we attempted to balance the number of social workers from within the main areas of social work (children's, adults' and mental health services) in order to follow principles 1 and 3 (2.1).

Our initial sampling plan was to collect all texts written by 50 social worker participants over a period of 20 days. We asked participants to keep a log of their writing over this timespan with the aim that all texts mentioned in the logs would be extracted, anonymised and shared with the research team, thus giving us a snapshot of texts produced over 1000 social worker days. However, a number of difficulties arose: 1) In order to secure sufficient participation and text collection, we involved additional LAs; 2) Whilst all social work participants (n = 71) were happy to be interviewed, permission was not given by all LAs to access their written texts; 3) A total of 29 social workers felt they had time available to keep a log of their writing; 4) Writing logs were not always kept over 20 consecutive working days due to other time commitments, sick leave and holidays.

Given the challenges involved in accessing such sensitive data, it is perhaps inevitable that our initial sampling strategy had to be rethought. Local Authorities have a clear duty of care towards children and adults using care services, with legal and ethical responsibility for protecting the personal data of vulnerable people, and we realised we could not fulfil Jaworska and Kinloch's idealised scenario of including 'all possible [textual] data produced in a given context' (Jaworska and Kinloch, 2018:114). We thus decided to adopt an opportunistic sampling approach, and - in order to fulfil our research council commitment to a 1-million-word corpus - accept all texts offered to us from participating social workers as far as was practical and ethical (see also discussion here: www.writinginsocialwork.com).

The resulting WiSP corpus includes a wide range of text types from mental health assessments to case notes relating to vulnerable children to emails between colleagues.[2] Corpus texts are particularly diverse in terms of length, varying from the very short (e.g. a 1-word email and 4-word casenote) to the very long (a 10,000-word court report). The corpus has differing numbers of texts by text category, social work domain and individual writer. Thus in some important ways there are limitations with the WiSP corpus from a corpus linguistics perspective in terms of representing the population of texts and controlling the wordcount (see 2.1, principles 1 and 3). However, the WiSP corpus remains the only corpus of social workers' writing currently available, and one of few corpora representing both sensitive and hard-to-access texts in any field (Leedham, Lillis and Twiner, 2020).

Following ethical approval and extensive consultation with LA gatekeepers, texts were collected and anonymised on the three sites before the research team were allowed access (3.2).

3.2. Building the corpus 1: anonymisation coding

As a condition of access to the texts, removal of all personal data from WiSP texts was required by participating authorities before handover to the research team. This removal was conducted by several individuals (two administrators in LA1 and one in LA 3, and two social workers in LA2). In order to remove personal data whilst maintaining core information, individual social care administrators - henceforward referred to as 'coders' - were provided with a set of anonymisation codes by the research team which was revised following discussion; e.g. service user [SU], the name of a service user's husband became [SUH], and a school name was coded as [SCHOOL]. Initially it was agreed that coders would include detail on the precise relationship, and on how many individuals were involved in a case e.g. coding the same person as [PERSON1] throughout a text. Such texts involve many participants – with one text including from [PERSON1] to [PERSON38] in a single text. This practice of marking individuals in a text is valuable as it provides an account of the complex network of people involved. However, the level of coding detail was gradually reduced due to constraints of coders' time, as well as project finances and the final deadline.

Initially, coders replaced direct identifiers such as names and locations, choosing to either manually replace words with codes, or use the 'find and replace' feature in MS Word. The latter was occasionally problematic as, for example, replacing all instances of the name 'Reg' with [PERSON] also changed 'regarding' to '[PERSON5]arding', and similarly 'Aila' changed 'available' to 'av[PERSON2]ble'. In each case, this rendered the individual service user's name recoverable from the text.

In contrast, Fig. 2 illustrates how coding has the benefit of maintaining contextual information as names are replaced with relationship codes, thus rendering the text easier to understand for an outsider. Methodologically therefore (and separate from the institutional requirement) having coders situated within the LA is advantageous.

In LA2, two social worker coders within children's care anonymised their own and other social workers' texts. While we had suggested codes such as [SUM], [SUH] to indicate 'service user mother/husband', the coders also wished to code for birth father/mother, prospective carer, foster carer, and so on (Fig. 3).

[1] The WiSP corpus is archived in the UK Data Archive. Access to the corpus is 'safeguarded' and requires an email request to the WiSP team. ReShare record 10.5255/UKDA-SN-853522.

[2] Nomenclature for text types are those supplied by social workers based on institutional labelling practices.

I explained that [OT] on Ward has highlighted further concerns about [SU]'s safety at home.

[SUD] said that she visits her mother almost daily, […]

[SUD] also explained that she feels that [SU] is now seeing things and having hallucinations. She remains low in mood with high anxiety - thinking that men are 'chasing her', and that she continues to stay the 'strangest things'.

Fig. 2. Extract from casenote: WISP1755 (OT is occupational therapist. SUD is service user's daughter).

[SW] Phone call to [MOTHERS-EX-PARTNER] (father to [HALF-SIBLING_2]). I spoke to [MOTHERS-EX-PARTNER] about the issues [service user-MOTHER] had raised regarding her contact with [HALF-SIBLING_2].

Fig. 3. Extract from casenote: WISP1215.

We are back in Court on [date] and the plan is to make a [issue] application. If it is granted [child] will be placed with [pc] with strict [issues] eg [issues] and more. Can we hold on till then?

Kind regards

Fig. 4. Extract from email: WISP3049.

[pc] works [time] daily, [activity] so Im not sure why she would be unable to see me on Friday during the day when all our meetings have taken place on weekdays?

Fig. 5. Extract from email: WISP0223.

[child] was asked how [child] felt living on [child] own; [child] said [child] was fine. I expressed concern that [child] was living on [child] own without someone supporting [child]. [child] said [child] had to start learning to live independently and was ok with it.

Fig. 6. Extract from casenote: WISP4206.

On occasion extensive coding of potentially sensitive areas was employed. This resulted in the content of some short texts becoming difficult to understand (Fig. 4).

The nature of the 'issue' in Fig. 4 is unclear; in particular, it is unclear which 'issues' are repetitions and which are newly mentioned. Similarly, in Fig. 5, information on the prospective carer's number of working hours and job has been removed by the social worker coder due to concerns over identifiability.

In Fig. 6, the coder attempted to render the text gender-neutral by removing mentions of the child's pronouns.

Fig. 6 illustrates the careful attention shown by social worker-coders in ensuring anonymity for individuals they have worked closely with which went beyond the institutional requirement to anonymise personal data.

The examples and discussion in this section illustrate how anonymisation coding is far from straightforward: considerable negotiation and discussion took place between members of the research team and 'coders' in the LAs, as well as coders' tacit knowledge coming into play in making decisions as to which information points are sensitive. The removal of detailed, inter-related personal information is a far greater task than the removal of personal data. Coders had different perspectives and reasoning affecting how they coded the data, with social worker coders in particular reflecting on their own practices when coding their documents. Furthermore, coding took place alongside text collection, meaning that codes evolved over time as new situations or relationships were encountered in the texts.

A further layer to ensuring satisfactory anonymisation from the perspective of social worker participants was the agreement that researchers would send draft papers and request additional checks on the anonymisation of data used in order to limit the risks of identifying any individuals through the 'jigsaw' of information from various sources (2.3). This commitment - which can be sustained where there is direct contact between researchers and participants - means that Corpus 1 is never fully abstracted from its context of production.

3.3. Building the corpus 2: archiving

As our commitment to the ESRC included full archiving of the WiSP corpus, the corpus coding needed to be both comprehensible and comprehensive for future unknown researchers. Following the initial coding within LAs (3.2), as a research team we discussed what further anonymisation was needed to render 'jigsaw identification' unlikely (2.3). For the

WiSP dataset overall, it seemed possible for an individual social worker or person written about to be identified simply through the sheer number of related data items (e.g. corpus texts, interview transcripts, fieldnotes); this concern was initially brought to our attention by social workers.

Within our research team of three, we held differing views over the extent to which the data should be anonymised for archiving. After extensive discussion, we decided to create two versions of the corpus: a full version for the research team's use which included engagement with participants about the use of any data extracts, and a version for archiving which was effectively a stand-alone corpus (henceforth respectively 'Corpus 1′ and 'Corpus 2′).

In preparing the two corpora we found (Carusi, 2008; Carusi and Jirotka, 2009) distinction between 'thin' and 'thick' identity helpful: whereas the former refers to proper names or other details which may identify a particular individual (mapping on to 'personal data'), the latter covers the 'narratives that people use to make sense of their lives and circumstances' (p.41) and aligns with our definition of sensitive data. Key questions were: would a corpus user identify anyone from the remaining information? E.g. would it be possible to cross check date and reason for hospital admission against another dataset? Would details appear in an online search covering open access articles quoting the extract? A fundamental concern for us as researchers was that the research should not have a negative impact on any individual's life, that is, to act with integrity as 'virtuous researchers' (e.g. Iphofen et al., 2017). We could envisage hypothetical situations where, for example, a care-leaver accessed the archive and was disturbed to recognise an account of a critical period of their life, or perhaps a researcher recognised a case history of someone they knew (cf. Tilley and Woodthorpe, 2011).

The impact of having two versions of the WiSP corpus is explored in Section 5. Here we outline the additional modifications applied to Corpus 2, the corpus made available for archiving, to decrease the risk of re-identification of individuals.

1) Extend the range of codes

As discussed in 3.2 the first stage of anonymisation in compiling the WiSP corpora was to replace direct identifiers such as personal names as well as spatial and temporal identifiers such as location and significant dates (e.g. of a court hearing). However, based on discussion with participating social workers and on our growing familiarisation with the data, we felt this was insufficient protection for individuals. For example a chronology covering a critical incident involving hospitalisation and police intervention was fully described stated with details of dates and drug doses. Coding in building Corpus 2 was thus extended to cover indirect identifiers such as drug names. Additionally, all dates were encoded as [DATE] since these form clear anchoring points which could be used to triangulate with databases containing court reports, police reports, hospital admissions or with social media. Each text was manually checked for any remaining personal identifiers. We also uploaded the corpus to the semantic tagging tool WMatrix (Rayson 2009) and checked the semantic domain of 'people' for omitted names. This detailed coding process was only possible due to the relatively small size of the corpus (1 million words).[3]

2) Delink texts

In preparing Corpus 2, it was agreed that 'delinking' (removing a connection between data items and information about data items) was important at several levels. First, we broke the link between individual texts and their writers (and by extension particular cases) by automatically renaming each text with a randomly-generated 4-digit number, rendering each text 'standalone'. Next, we removed coding which linked social worker and service user. Contextual data on individual social workers such as age group and length of service was not provided within the metadata as felt this better fulfilled the spirit of our agreement with LAs.

3) Remove highly sensitive individual texts

A number of social workers were concerned that the vulnerable adults and children mentioned in texts might remain identifiable post-anonymisation due to their unique situation, family or medical issues (cf. discussion in Perera et al., 2016). We thus offered all participants the option to mark texts as 'not for sharing' (NFS) at the point of submission if they wished them to be used by the project team only (and thus constituting part of Corpus 1) and not archived (in Corpus 2).

In WiSP, some participating social workers selected particular texts to be included in the project, and excluded others based on their perspectives of the level of sensitivity, whereas other participants granted access to all texts they had written within the time period of the study. This difference foregrounded the issue of who should decide whether the risks are too great to share a text: the participating social workers or the research team. Following extensive team discussion, a number of key principles were agreed (3.4) to underpin decisions around which additional texts should be removed from the archived corpus (Corpus 2).

4) Restrict users (environment)

Within the UK Data Archive we chose the most controlled data environment, requiring researchers to register and also contact the WiSP team to request access from the team.

While the removal of data as described in 1–3 above goes against principle 4 of corpus-building (3.2), we felt that reducing the links between individuals and the texts written about them was necessary to adhere to the ethical principle of doing no harm and thus retain researcher integrity (Iphofen et al., 2017).

3.4. Comparison of corpus 1 and corpus 2

The difference between Corpus 1 and Corpus 2 is relatively slight at just 38 additional texts in the former (Table 1), comprising approximately ten per cent of the wordcount of Corpus 1.[4] Most omitted texts are assessment reports (19) and casenotes (17) as these are generally longer, averaging 2–3000 words, and more detailed than emails, letters and memos (the remaining 2 are chronologies).

Of the excluded texts, six were designated by social workers as NFS beyond the research team due to their highly sensitive nature (see 2.2). A further 32 texts were excluded by the research team, largely on the grounds of multiple reference points to an individual's life (e.g. a 10,000 word chronology). Our rationale for excluding texts comprised combinations of the following:

a) Extensive detail on personal, medical and/or family history.
b) Detail of other restrictions in place e.g. 'abduction order', 'contact order'.
c) Details of identifying marks, height, weight, IQ, injuries, illnesses, medical issues/appointments/investigations, medications, therapies.

Table 1
Comparison of the two WiSP corpora.

Corpus 1		Corpus 2	
Tokens	*No. texts*	*Tokens*	*No. texts*
1003,096	4608	966,238	4570

[3] Since WiSP was archived, more support with coding and pseudonymisation is now available on the UKDA site via a text anonymisation helper tool.

[4] This difference in size of Corpus 1 and Corpus 2 is not fully reflected in Table 1 due to additional anonymisation coding in Corpus 2. Wordcounts are from WordSmith Tools v.7 (Scott 2019).

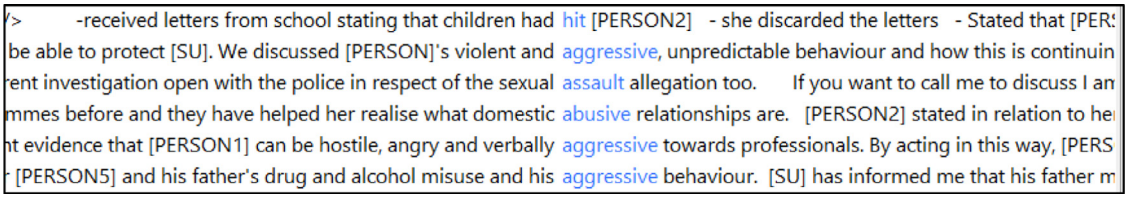

Fig. 7. Example concordance lines for semantic area: E3- VIOLENT/ANGRY.

d) Extensive detail around family relationships (e.g. siblings/half/step siblings; partners; grandparents).
e) Descriptions of behaviours and patterns of interactions.
f) Details of individuals' awards/achievements.
g) Extensive use of direct quotes.
h) Extensive repetition across texts.
i) Sets of texts around a single case.

The next section explores the extent to which differences in the two corpora affect understandings about the nature of social worker writing.

4. Exploring the WiSP corpora

This section explores the possible epistemological implications arising from the existence of two corpora using the WiSP corpora as an illustrative case study, focusing on three worked examples. The first employs the whole of each corpus, comparing findings from Corpus 1 with Corpus 2; Example 2 slices the corpora by text category and Example 3 returns to the whole corpora for analysis of one lexical item. Extracts included in this section from 'NFS', following our agreement with participants, have undergone further minor wording changes and additionally have been confirmed by participants as sufficiently anonymised. In each worked example, the wider co-text is examined to explore the context of each concordance line or keyword.

4.1. Worked example 1: using semantic tagging to unveil the broad 'preoccupation' of social work writing

The first worked example uses WMatrix semantic tagging software to uncover similarities and differences between Corpus 1 and Corpus 2, when compared to a reference corpus (see also Leedham, Lillis and Twiner, 2020). Wmatrix (Rayson, 2008) was used to extract key lexical items from each corpus in turn using British English 2006 (BE06), a 1-million-word corpus of published general written British English (Baker, 2009), as a reference corpus.[5]

Comparison with BE06 indicates that Corpus 1 and Corpus 2 contain broadly equivalent key semantic categories, including the domains of SOCIAL ACTIONS, STATES AND PROCESSES; HEALTH AND DISEASE; MEDICINES AND MEDICAL TREATMENT; WORRY; PEOPLE; HELPING; SPEECH: COMMUNICATIVE; TIME; NEGATIVE; and WORK AND EMPLOYMENT. These semantic areas indicate the 'preoccupation' (Baker 2010) of social work writing (see also Leedham, Lillis and Twiner, 2020). This similarity is of course to be expected, since the two corpora comprise largely the same texts.

However, several *additional* semantic areas were found to be key in Corpus 1 when compared to BE06 but were not key areas in Corpus 2. These categories include VIOLENT/ANGRY; SAD; SMOKING AND NON-MEDICAL DRUGS; DAMAGING AND DESTROYING; CRIME. Example concordance lines for two semantic areas are presented in Figs. 7-8.

Much of the lexis within these semantic domains also occurs in Corpus 2, though with the removal of the NFS texts does not reach the specified cut-offs; this discrepancy between the corpora suggests that the 38 omitted texts contain details of more extreme cases involving abuse, drugs and crime. While both corpora reveal the same broad semantic fields, research using the archived corpus would not uncover the more violent and highly challenging contexts that social workers engage in, meaning that Corpus 1 makes visible a broader area of social worker 'preoccupation' (Baker, 2010) than Corpus 2.

4.2. Worked example 2: use of quotations in assessment reports

Worked Example 2 focuses on the text category of assessment reports and limits analysis to a single rhetorical feature: the use of quotations. The broad category of assessment reports covers a range of documents such as parenting plans, placement plans and court reports. This section first provides quantitative findings, then explores the functions of 200 quotations.

We used AntConc (Anthony 2008) to search for a character space followed by single or double quotes (avoiding possessive or contracted *'s*) and checked to exclude quotations within form language. This method produced 634 hits in Corpus 1 reports (244 files) and 510 hits in Corpus 2 reports (227 files), indicating that usage is higher in Corpus 1 to a statistically significant level (LL7.44,% DIFF 17.57). It is important to note that widespread use of direct quotes formed one consideration in the research team's decision to remove texts from Corpus 1 (3.4); this resulted in an additional 17 assessment reports in Corpus 1.

A randomly-extracted set of 200 concordance lines containing quotations were examined from Corpus 2 reports by one researcher with 10% of these checked by a second researcher. This analysis suggests that smaller categories include scare quotes (7%, $n = 14$), citations of other professionals (health, school, police) (<6%, $n = 11$), titles of books/websites (<3%, $n = 5$), and quotes from own writing or from other social workers (<2%, $n = 3$). The remaining 82% (163) of the sampled concordance lines comprise quotations attributed to service users or their relatives/friends, usually in the form of direct quotations (NB 2% ($n = 4$) are text messages).

Use of quotations are either utterances presented as being witnessed directly by the social worker (Example 1) or those from a third party reported by a service user or relative to the social worker (2).

1) When asked direct why he thought he was in hospital he answered in one word syllables "hygiene", "unclean" "smell" "shower". (WISP2282, mental health).
2) [PERSON1] said that [PERSON2] mother is 'amazing'. She is supportive and offers practical support including transport to necessary appointments, (WISP0534, Children's).

Direct quotations of service users' language, embedded within social worker accounts, may be used to indicate strength of feeling (3) or serve as evidence for the social worker's analysis (4).

3) [child2] was asking [child] to stop it. [child] told him to "fuck off". (WISP1761, children's).
4) However, given the lifestyle she returned to when in [LARGE-CITY], [pc] feel that [bm] must have 'played' and manipulated them. (WISP0583, Children's).

In Corpus 1, the proportions are similar, with a greater weighting towards service user/relatives' language due to the NFS texts. Quotations

[5] Log Likelihood was set to 0.01% level; $p < 0.0001$; critical value = 15.13, minimum frequency was set to 50; effect size measure used was %DIFF.

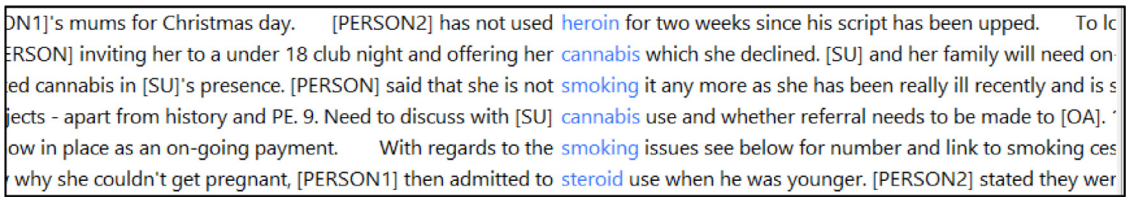

Fig. 8. Example concordance lines for semantic area: F3 SMOKING AND NON-MEDICAL DRUGS: BAD.

from NFS texts contain more swearing and also accounts of concerning incidents (5, 6).

5) She informed [PERSON] that [PERSON2] got "really drunk" last night, "smashed up stuff in the house the house, took all [PERSON1]'s money, got in the car whilst drunk (Corpus 1, chronology)
6) [SU1] stated with the direct work that his "daddy was nasty and he needs to be in a nasty house". (Corpus 1, assessment report)

Thus while the two corpora again provide similar findings, the addition of the Corpus 1 NFS texts containing more serious or extreme cases enables greater analysis of critical incidents within assessments (see also Leedham, 2020).

4.3. Worked example 3: exploring a lexical item

This worked example focuses on the lexical item *asleep*, selected as an illustrative example of a seemingly 'neutral' word but one which is frequently used to indicate an evaluative stance. *Asleep* occurs 53 times within Corpus 1 and 49 times in Corpus 2. We begin the analysis with the archived Corpus 2.

In Corpus 2, a total of 19 occurrences of *asleep* describe children. The majority of these (13) comment neutrally on a child being asleep (Example 7):

7) [SU] was tired and she fell **asleep** in the car. (WISP2868, casenote [bold added])

Four instances comment positively on a young child being asleep. In (8), the social worker comments on sleep to illustrate the warm relationship between child and carer.

8) [CHILD] looked comfortable with [PERSON] and he fell **asleep** on her. He appeared content in the home and [PERSON] was attentive to his needs. (WISP1643, casenote).

Just two instances are negative: in (9) commenting on falling asleep is given within a list of factors describing the home situation as chaotic.

9) She is not keeping him clean, there was no fresh food at home. His bed was not made and all rooms were very dirty, untidy and smelly. [CHILD] said he sometimes slept on the sofa with Mum and fell **asleep** watching TV with her. The home conditions and parenting are not good enough at present. (WISP2066, casenote).

A second negative instance describes a child napping for a long time during an unannounced visit.

In contrast, commenting that an adult is asleep (n = 30) is mainly negative with just two positive instances (describing falling asleep at the end of a 'perfect day') and two neutral. Most examples are from the domain of children's care and describe adult carers being asleep and missing a case conference, falling asleep on the sofa at night or a child being unable to wake their parent. In (10), the parent has apparently been asleep prior to the social worker visit:

10) We knocked at the door several times before we got an answer. [PERSON] then came to the door and appeared very sleepy as though he had just been woken up. [CHILD] was wondering about and appeared happy. [PERSON] invited us in. I asked [PERSON] if he had been **asleep** and he stated he was very tired from doing his night shift, he stated he hadn't been **asleep** but his eyes had been shutting but he could still hear everything. (WISP3610, casenote)

The account of the father's reported delay in answering the door and implication that he had just woken up implies concern on the part of the social worker about whether adequate care is being provided for the child.

Daytime sleeping is also commented on within a case of suspected substance abuse as evidence of a chaotic lifestyle:

11) During a conversation with [PERSON6] today she mentioned that she'd spoken to [PERSON7] (brother) over the weekend who told [PERSON6] that he knew [PERSON2] was using heroin again and stealing, that he was **asleep** most the day and out most the night. (WISP1414, casenote, children's).

For older adults under social worker care, being asleep during the day can be a sign of deterioration in health (12,13).

12) … [PERSON2] stated that [SU] was just sat in the lounge **asleep** when she arrived and he didn't really speak with her. (WISP3455, casenote)
13 [SU] still unsteady mobilising and spending a lot of time **asleep**. (WISP0086, casenote)

Exploration of concordance lines in Corpus 2 suggests that, while at face value a 'neutral' term, the decision to comment on a child, adult carer or adult service user being asleep is often a marked choice. While commenting on a young child sleeping is predominantly neutral or viewed positively as evidence of a strong, caring bond between adult and child, commenting on an adult sleeping is often evidence of inadequate care-giving or, for older adults, of deterioration in health. Corpus 1 contains just four additional occurrences of *asleep*, all of which are classed as negative (14,15):

14) [SU1] fell **asleep** within lesson after lunch break and [PERSON1] question whether this was due to him smoking cannabis (Corpus 1, assessment report)
15) [CHILD] stated he started a paper round to get extra money for himself however [CHILD] would fall **asleep** at school. (Corpus 1, casenote)

The examples suggest that sleeping in the day is viewed by the social worker as possible evidence of cannabis usage (14) and that the child is not able to cope with their daily routine (15). The additional examples from Corpus 1 thus broaden the range of inferences accorded to evaluation through commenting on sleep by adding the example of an older child, but do not fundamentally alter the analysis and findings that are possible through Corpus 2 alone.

4.4. Summary

From the three worked examples it can be seen that the two WiSP corpora provide the same broad findings, as would be expected. However, the additional texts in Corpus 1 contain more sensitive material – the reason for their exclusion – and analysis of Corpus 1 therefore uncovers a greater range of semantic areas. Omitted texts are likely to be the

more sensitive reports, meaning the archived corpus contains a higher proportion of less contentious or 'safer' texts. In terms of the overall purpose of building a WiSP corpus of written texts - to characterise the nature of professional social work written discourse - Corpus 1 enables a more comprehensive account, foregrounding the often profoundly challenging situations social workers are writing about and indicating some of the complexities of capturing these in written discourse. Corpus 2 necessarily offers a more restricted view, potentially minimising such complexity.

5. Conclusion

In this paper we set out to answer three questions. Here, we summarise findings relating to these questions based on our work with the WiSP corpus.

RQ1 What are the challenges in data preparation for archiving hard-to-access, sensitive textual data, particularly around anonymisation coding?

The initial challenge in compiling the WiSP corpus was to gain access to LAs and to negotiate the involvement of administrators and social workers in each site to fulfil the institutional requirement of removing all personal data before texts could be made available to researchers. A further challenge was developing an anonymisation coding scheme to ensure that important contextual detail was maintained whilst protecting identities of individuals. The final challenge was to meet the expectations of participating social workers – as well as to ensure we were upholding ethical integrity as researchers – in ensuring jigsaw identification was unlikely; this went beyond the institutional requirement to remove personal data. In order to meet these expectations and due to our reaching a deeper understanding as to what might be at stake for people mentioned in the texts, a second corpus was created.

RQ2: Given the adopted solution of creating a separate corpus for archiving, how do the differences between the two versions of the WiSP corpus affect findings?

Based on analysis of three worked examples, it is clear that while the two corpora largely comprise the same texts, the additional texts in Corpus 1 give extra perspectives. The examples presented in Section 4 illustrate that useful analyses can still be conducted on both Corpus 1 and Corpus 2 datasets, with real-world implications for the understanding of social worker writing. Corpus 1 extends what it is possible to reveal about social work writing as it contains more texts of 'extreme' sensitivity. In addition, Corpus 1 texts contain more metadata on social work writers and the texts produced, meaning that it is possible to combine corpus analysis with examination of other WiSP datasets on the same case.

RQ3: To what extent do the issues raised and solutions found problematise conventionally-accepted principles in corpus building?

The decisions taken in building the WiSP corpus raise important questions about key principles within corpus linguistics. The first set of questions relates to the building of what is referred to throughout as Corpus 1. A number of decisions were made which do not align with these principles - such as accepting all available texts and removing metadata - but which were considered essential in order to build a corpus of sensitive and hard-to-access texts. Whilst not reflecting certain core principles - in particular those signalling a more positivist orientation to knowledge making such as generalisability - we consider that there is substantial value in making such compromises as corpus linguistic tools can generate insights into textual practices which might otherwise not come (easily) to light.

The second set of questions relates to the decisions to make a slightly modified version available for secondary analysis, referred to throughout as Corpus 2. The common funding requirement to make corpora available raises fundamental ethical concerns about data usage: whilst additional checks with participants can be facilitated by the immediate research team, no such checks are possible once the data is abstracted from context, thus leading to the decision in the case of WiSP to build two slightly different corpora. The existence of two versions is unusual in corpus linguistics but we think is valid for the ethical reasons discussed in 3.3.

Due to the varied texts, the lack of sampling frame and its relatively small size, we cannot claim the WiSP corpus is wholly representative of social workers' writing, as is often a stated goal in corpus linguistics (2.1). As with all small corpora, it is particularly important to bear in mind the make-up of the datasets in terms of number of writers (38), LAs (3) and proportions within each domain in order to avoid any danger of over-claiming from this sample to the larger population. We are also acutely aware that the steps taken to render re-identification extremely unlikely involve data loss in terms of reduced textual metadata and the delinking of texts (see also Tucker et al., 2016). In terms of our responsibility to participants, authorities and those being written about, ethical considerations have to take priority. Instead, the corpus can be viewed as a 'way in' to discourse analysis, using the methodology of corpus-assisted discourse studies (Taylor and Marchi 2018). While in corpus terms, the WiSP corpora are perhaps more imperfectly formed than corpora comprising readily-attainable texts, yet as a rare corpus of hard-to-access texts from a particularly sensitive area the production of such a corpus at all is a valuable contribution. An important overall learning point from our close work with text *creators* is that social workers' concerns ensured that as text *researchers* we learned to treat the texts with care, as living data with consequences for individuals (Lillis, Leedham and Twiner, 2017) rather than purely as data items abstracted from producers, receivers, social practices and consequences.

The WiSP corpus is to date the only available collection of social worker writing and represents a significant first step in corpus creation for social work writing, extending the range of texts available to researchers and enabling more systematic text-based research into this professional discourse than has previously been possible. We envisage that Corpus 2 will lead the way for future researchers of social work writing practices and will also be useful to social work educators: the texts have already been useful in the creation of a site for trainers and students (WiSP). We anticipate that the discussions and potential solutions around anonymisation issues presented in this paper will be beneficial for future corpus builders, particularly in the case of sensitive datasets.

Funding

The WiSP research project was funded by the Economic and Social Research Council ES/M008703/1 and The Open University, UK. The research was carried out by Theresa Lillis (PI), Maria Leedham (Co-I) and Alison Twiner (Research Associate). We would like to thank the participating local authorities and the social workers who so generously took part in the WiSP project but who have to remain anonymous for confidentiality reasons.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We would like to thank Dana Therova for her work supporting the literature review for this paper.

References

Anthony, Laurence. 2008. 'Antconc'. http://www.antlab.sci.waseda.ac.jp/software.html.
Baker, Paul. 2010. *Sociolinguistics and Corpus Linguistics*Edinburgh University Press.
Baker, Paul, 2018. Conclusion: reflecting on reflective research. In: Taylor, Charlotte, Marchi, Anna (Eds.), Corpus Approaches to discourse: A critical Review. Routledge, London and New York.
Biber, Douglas., 1993. Using register-diversified corpora for general language studies. Computational Linguistics 19, 219–242.

Bolly, Catherine T., Boutet, Dominique, 2018. The multimodal CorpAGEst corpus: keeping an eye on pragmatic competence in later life. Corpora 13, 279–317.
Carusi, Annamaria., 2008. Data as representation: beyond Anonymity in e–research ethics. Int. J. Internet Research Ethics 1, 37–65.
Carusi, Annamaria, Jirotka, Marina, 2009. From data archive to ethical labyrinth. Qualitative Res. 9, 285–298.
Elliot, Mark, Mackey, Elaine, O'Hara, Kieron, 2020. The Anonymisation Decision Making Framework: European Practitioners' Guide. UK Anonymisation Network, Manchester.
Elliot, Mark, Elaine Mackey, Kieron O'Hara, and Caroline Tudor. 2016. "The Anonymisation Decision-Making Framework: UKAN." In. Manchester.
Hayes, David, Devaney, John, 2004. Accessing Social Work Case Files for Research Purposes:some Issues and Problems. Qual. Soc. Work 3, 313–333.
Iphofen, Ron, Robert Dingwall, Janet Lewis, John Oates, and Nathan Emmerich. 2017. 'Towards Common Principles for Social Science Research Ethics: a Discussion Document for the Academy of Social Sciences.' in.
Irwin, S., 2013. Qualitative secondary data analysis: ethics, epistemology and context. Prog. Develop. Stud. 13, 295–306.
Jaworska, Sylvia, Kinloch, Karen, 2018. Using multiple data sets. In: Taylor, Charlotte, Marchi, Anna (Eds.), Corpus Approaches to discourse: A critical Review. Routledge, London and New York.
Koester, Almut., 2010. Building small specialised corpora. In: O'Keeffe, A., McCarthy, Michael (Eds.), The Routledge handbook of Corpus Linguistics. Routledge, London & New York.
Kreischer, Kim-Sue., 2019. The relation and function of discourses: a corpus-cognitive analysis of the Irish abortion debate. Corpora 14, 105–130.
Lischinsky, Alon., 2018. Overlooked text types: from fictional texts to real-world discourses. In: Taylor, Charlotte, Marchi, Anna (Eds.), Corpus Approaches to discourse: A critical Review. Routledge, London and New York.
Leedham, M. (2021). 'Social Workers Failed to Heed Warnings': A Text-Based Study of How a Profession is Portrayed in UK Newspapers', The British Journal of Social Work. https://doi.org/10.1093/bjsw/bcab096.
Leedham, M., Lillis, T., Twiner, A. (2020). Exploring the core 'preoccupation' of social work writing: A corpus-assisted discourse study. Journal of Corpora and Discourse Studies. 3. Pp.1–26. https://jcads.cardiffuniversitypress.org/articles/abstract/26/.
Lillis, T., Leedham, M. and Twiner, A. (2019). "Writing in social work professional practice (2014–2018)." In. Colchester, Essex: UK Data Service. 10.5255/UKDA-SN-853522.
Lillis, T., Leedham, M., Twiner, A. (2020). Time, the written record, and professional practice: The case of contemporary social work, Written Communication 37, 4. https://doi.org/10.1177/0741088320938804.
Lillis, T., Leedham, M. and Twiner, A. (2017, 2020). 'If it's not written down it didn't happen': Contemporary social work as a writing intensive profession. Journal of Applied Linguistics and Professional Practice.14,1. Pp.29–52. https://doi.org/10.1558/jalpp.36377.
Marchi, Anna, Taylor, Charlotte, 2018. Introduction: partiality and reflexivity. In: Charlotte, Taylor, Marchi, Anna (Eds.), Corpus Approaches to discourse: A critical Review. Routledge, London and New York.
McEnery, Anthony, Hardie, Andrew, 2012. Corpus Linguistics. Cambridge University Press, Cambridge, UK.
Nesi, Hilary, Gardner, Sheena, 2018. The BAWE corpus and genre families classification of assessed student writing. Assessing Writing 38, 51–55.
Ohashi, Yukiko, Katagiri, Noriaki, Oka, Katsutoshi, Hanada, Michiko, 2020. ESP corpus design: compilation of the Veterinary Nursing Medical Chart Corpus and the Veterinary Nursing Wordlist. Corpora 15, 125–140.
Parry, O., Mauthner, N., 2004. Whose data are they anyway?: Practical, legal and ethical issues in archiving qualitative research data. Sociology 38, 139–152.
Perera, Gayan, Broadbent, Matthew, Callard, Felicity, Chang, Chin-Kuo, Downs, Johnny, Dutta, Rina, Fernandes, Andrea, Hayes, Richard D, Henderson, Max, Jackson, Richard, Jewell, Amelia, Kadra, Giouliana, Little, Ryan, Pritchard, Megan, Shetty, Hitesh, Tulloch, Alex, Stewart, Robert, 2016. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. BMJ Open 6, e008721.
Rayson, Paul. 2009. "Wmatrix: a web-based corpus processing environment." In. Lancaster: computing Department, Lancaster University.
Scott, Mike. 2019. 'WordSmith Tools, v.8', Lexical Analysis Software. http://www.lexically.net/wordsmith/version5/index.html.
Sinclair, John. 2005. 'Corpus and Text: basic Priniciples.' in Martin Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*.
Sinclair, John M., 1991. Corpus Concordance Collocation. Oxford University Press, Oxford.
Staples, Jennifer M., Bird, Elizabeth R., Masters, Tatiana N., George, William H., 2018. Considerations for Culturally Sensitive Research With Transgender Adults: a Qualitative Analysis. J Sex Res 55, 1065–1076.
Swales, J.M., 2004. Research Genres: Explorations and Applications. Cambridge University Press, Cambridge, UK New York.
Taylor, Charlotte, and Anna Marchi. 2018. *Corpus approaches to discourse: a critical review*Routledge: London and New York.
Tilley, Liz, Woodthorpe, Kate, 2011. Is it the end for anonymity as we know it? A critical examination of the ethical principle of anonymity in the context of 21st century demands on the qualitative researcher'. Qualitative Research 11, 197–212.
Tognini-Bonelli, E., 2001. Corpus Linguistics at Work. John Benjamins, Amsterdam.
Tucker, Katherine, Branson, Janice, Dilleen, Maria, Hollis, Sally, Loughlin, Paul, Nixon, Mark J., Williams, Zoë, 2016. Protecting patient privacy when sharing patient-level data from clinical trials. BMC Med. Res. Methodol. 16, 77.
UK Government. 2018. 'General Data Protection Regulation (GDPR)', Accessed 09/06/2021. https://www.gov.uk/data-protection.